Measuring Numeracy: Validity and the Programme for the International Assessment of Adult Competencies (PIAAC)

Tunstall, Samuel L.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Tunstall, S. L. (2020). Measuring Numeracy: Validity and the Programme for the International Assessment of Adult Competencies (PIAAC). *Numeracy*, *13*(2), 1-37. https://doi.org/10.5038/1936-4660.13.2.1348

Numeracy
Advancing Education in Quantitative Literacy

Volume 13 | Issue 2 Article 6

2020

Measuring Numeracy: Validity and the Programme for the International Assessment of Adult Competencies (PIAAC)

Samuel L. Tunstall
Trinity University, stunstal@trinity.edu

Recommended Citation

Tunstall, Samuel L.. "Measuring Numeracy: Validity and the Programme for the International Assessment of Adult Competencies (PIAAC)." *Numeracy* 13, Iss. 2 (2020): Article 6. DOI: https://doi.org/10.5038/1936-4660.13.2.1348

Measuring Numeracy: Validity and the Programme for the International Assessment of Adult Competencies (PIAAC)

Abstract

A tension raised in recent scholarship is that between numeracy as a social practice and numeracy as a functional skill set. Such frameworks for conceptualizing numeracy pose a challenge to assessment because what individuals do with numeracy is not the same as what individuals can do (or express) in an assessment setting. This study builds on work related to numeracy assessment through a validity examination of a portion of a well-known assessment: the OECD's Programme for the International Assessment of Adult Competencies (PIAAC). In following a path set out by standards for assessment, I ask: What does the PIAAC numeracy assessment claim to measure? What are the intended uses of the assessment? How are we to interpret scores with those uses in mind? And to what degree do evidence and theory support interpretations for those uses? The main finding from this work is that while score interpretations from the PIAAC numeracy assessment may be valid for the use of describing proficiency distributions for specific groups, the construct of interest—numerate behavior—is not what is measured. Moreover, evidence distinguishing what is measured from other constructs, such as the OECD's conception of literacy, is largely absent. This study contributes to existing literature on numeracy assessment by providing sources of evidence to consider in making judgments about validity for an assessment. It also suggests that, as scholars, we carefully hedge the ways that we talk about large-scale assessments, and in relation, what individuals can or cannot do based on results from such assessments.

Keywords

numeracy, quantitative literacy, quantitative reasoning, validity, assessment, social practices

Cover Page Footnote

Samuel Luke Tunstall is Director of the Quantitative Reasoning and Skills Center at Trinity University in San Antonio, Texas. He is currently Vice-President of the National Numeracy Network, and Chair of the Mathematical Association of America's Special Interest Group in Quantitative Literacy. His research interests include quantitative literacy practices and the assessment of quantitative literacy.

Introduction: The Social Context of This Investigation

Proceeding from a catchy title, "U.S. Millennials Post 'Abysmal' Scores in Tech Skills Test, Lag behind Foreign Peers," *Washington Post* columnist Frankel (2015) noted

> There was this test. And it was daunting. It was like the SAT or ACT—which many American millennials are no doubt familiar with, as they are on track to be the best educated generation in history—except this test was not about getting into college. This exam, given in 23 countries, assessed the thinking abilities and workplace skills of adults. It focused on literacy, math and technological problem-solving. The goal was to figure out how prepared people are to work in a complex, modern society. And U.S. millennials performed horribly.

Frankel is not the only journalist in popular media to participate in discussions about aggregate results of Americans' performances on international assessments. Similar headlines sounding nearly identical alarms about performance abound in relation to both this exam (e.g., Zinshteyn 2015; Emanuel 2016) and similar ones from the past (e.g., National Commission on Excellence in Education 1983; Rice 2009). In the particular piece excerpted above, Frankel discusses with an Educational Testing Service (ETS) researcher US millennials' results from the Programme for the International Assessment of Adult Competencies (PIAAC).

Developed by the Organisation for Economic Co-operation and Development (OECD), the PIAAC is a relative to an older (albeit still in use) OECD exam, the Programme for International Student Assessment (PISA). The PIAAC differs from PISA in that (among other things) the former is primarily aimed at individuals aged 16 to 65, rather than 15-year-olds—the sole group taking part in PISA. Individuals participate in PIAAC in their residences, rather than in school, as is the case with PISA. With data collection completed from 2011–2012, the first administration of PIAAC consisted of a survey of 166,000 adults aged 16 to 65 in twenty OECD member countries (in addition to Cyprus and the Russian Federation); the second administration is currently in progress. Per the OECD, PIAAC "assesses the proficiency of adults from age 16 onwards in literacy, numeracy and problem solving in technology-rich environments," the motivation being that such proficiencies "are relevant to adults in many social contexts and work situations, and necessary for fully integrating and participating in the labour market, education and training, and social and civic life" (2013b, 5). In addition to testing in literacy, numeracy, and problem-solving in technology-rich environments, respondents also complete a detailed questionnaire, which includes demographic information (e.g., the level of education of one's parents) as well as habits in relation to numeracy, literacy, and one's general home life.

The first paragraph of Frankel's article represents the PIAAC from a particular perspective, one that differs from my own in that, in my view, the PIAAC assessment

- is not necessarily daunting (the test lasts around 60 to 80 minutes, which includes time for the background survey),
- is not readily comparable to the SAT or ACT (the format, the constructs tested, and stakes for test takers are different),
- is taken by few Americans (5,010 people in the 2011–2012 administration), and
- aims to assess the construct of numeracy, rather than that of mathematics (which the test developers distinguish, as I discuss later).

It is not wholly surprising that my view of the PIAAC is different from that of Frankel, and my purpose here is not to admonish or belittle Frankel. Journalists often incorporate influences and perspectives that are different from mathematicians and research scientists when adapting research studies into news products suitable for their respective audiences (Woloshin and Schwartz 2002). Given the task that journalists face in translating complex ideas into bites accessible to a wide audience, it is understandable that these differences in perspective might arise. For example, US readers may not be familiar with the term *numeracy*, but they probably have some familiarity with the term *mathematics*. The substitution in terminology likely does little harm in that context. Indeed, it may be a necessary substitution for the work to be accessible to Frankel's readership. That being said, what I have found surprising, and what partially prompted the exploration I report on here, is the degree to which interpretations of PIAAC results by PIAAC researchers are valid for proposed uses by the assessment's developers.

That is, though I was not familiar with the concept at the time, I was concerned with the *validity* of the PIAAC numeracy assessment in the context of interpretations such as those from Frankel in the title and body of the article—a concern that is not completely new in the context of the PIAAC (e.g., Evans 2014; Oughton 2018). By validity, I mean the degree to which interpretations of scores are appropriate for their proposed uses. Are Americans, on the aggregate, actually unprepared to work in a "complex, modern society"?

Warrants for the Investigation

My rationale for this work stems from two areas: (1) my personal connection to coursework centered around numeracy, and (2) calls for increased interest in assessment as it relates to numeracy. With respect to the first area, my personal connection comes from teaching courses centered on quantitative literacy at both two- and four-year institutions. I write about this personal connection, or my positionality (Foote and Bartell 2011), because it inevitably informs the work that I do, regardless of whether I desire it. In my teaching, I have found that the ways

my students think about and approach real-world contexts often differs from how I pose or broach them in formal assignments like labs or quizzes. A recent example of this disconnect occurred in the 2018 Summer Session at Michigan State University (MSU), when I facilitated a unit on gerrymandering for a course I was teaching, Quantitative Literacy II (see Tunstall et al. [2016] for more information about these courses). A bulk of the unit was on the mathematics of the efficiency gap (Stephanopoulos and McGhee 2015), but that topic—even the YouTube video[1] associated with it—was not the first thing that arose in students' beginning-of-class discussions; instead, it was voter suppression and proportional representation, the former of which had been a hot topic in the news that month. To subsequently read Frankel's headline not long after those conversations, which suggests that Americans' numeracy scores are abysmal, yielded dissonance for me. I saw promise, not deficit, in students' discussions about voting and representation. Students were engaged with the material and ready to learn about the efficiency gap. Furthermore, my students were not answering the types of questions sampled in Frankel's article in class, and it was difficult to imagine them answering many of them in any current context—whether in or out of class. This raised a question: millennials performed poorly by what standards?

With respect to my second rationale for this exploration, note that US-based scholars of numeracy and quantitative literacy have expressed increased interest in assessment of numeracy, quantitative literacy, and quantitative reasoning in the last decade (Vacher 2015; Cahoon and Kiliç-Bahi 2019). This interest stems from larger movements to assess general education outcomes in higher education (Rhodes 2010), as well as the more specific need to gauge the success of novel programs in numeracy, where success is measured by the extent to which (in this case) college graduates are able to demonstrate behaviors and attitudes aligned with—that is, are valid proxies for—what has been defined as numeracy, or quantitative literacy. Outside of the US context, numeracy has been (and continues to be) studied by scholars in various communities, notably including the international forum Adults Learning Mathematics,[2] where PIAAC has been questioned and critiqued (Evans 2014), albeit not through the lens used here: that of a unitary concept of validity.

In relation to the aforementioned point, note that as scholars, our ability to make claims based on an assessment is contingent upon the validity (i.e., alignment of purposes) of that assessment (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). While some in the field have alluded to the importance of validity in developing assessments for numeracy (e.g., Gaze et al. 2014), and even referenced notions of validity in analyses of PIAAC (Evans 2014; Tout et al. 2017),

[1] See https://www.youtube.com/watch?v=IKtbfVmKM3w for the video from WNYC.
[2] See http://alm-online.net/ to learn more about the community, its annual conference, and related publications.

heretofore there has been no holistic consideration of the validity of a numeracy assessment—that is, the consideration of more than just one facet of validity (for example, in the case of Gaze et al. [2014], content validity). Until the mid- to late-twentieth century, validity was viewed through multiple lenses, or multiple types of validity. These types included (among others) content validity, criterion validity (consisting of predictive and concurrent validity), and construct validity. Insofar as validity is now viewed from a broader lens than just one of a specific type of validity, and the justifiable use of an assessment is contingent upon a foundation of validity (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014), this paper provides an example of what the validation process might look like as we consider the types of claims we can make from an assessment. This important consideration is the primary contribution of this article to the field of numeracy scholarship.

My work is informed by a social practices view of numeracy (Craig and Guzmán 2018; Oughton 2018) demonstrating that a social theory of numeracy need not be in opposition with epistemological expectations for rigor and method expected by many individuals in the educational research community (e.g., Shulman 1981; Scheaffer 2008). Working from these rationales, I embarked on a post hoc validity exploration of the numeracy portion of the PIAAC, using an argument-based approach to validation (Kane 2012; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). In particular, I discuss validity and the validation process from an external standpoint of the PIAAC, raising questions and considerations for developing, implementing, and reporting on their assessments related to numeracy. Further, to those ends, I begin by discussing a definition of validity, and then discuss assessments of numeracy. I then transition to the PIAAC, and a discussion of the validity of PIAAC interpretations in light of the test developers' proposed uses. I end with implications and a call for future work in relation to validation and numeracy assessments.

Definition of Concepts

Prior to exploring validity in relation to the PIAAC numeracy assessment, it is important to have a foundation for what validity is. I begin this section with that grounding discussion. The definition I adopt, and that I will explain in further detail below, is that validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of test scores" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014, 11).

As an adjective, *valid* is a relative term insofar as it raises questions of: Valid to whom? Valid with respect to what? And valid by what standard(s)? For example,

DOI: https://doi.org/10.5038/1936-4660.13.2.1348

the declaration, "She brings up a valid point," bears little meaning without knowing more about the conversants, the referent for any claim of validity, or the backdrop of their conversation. Even with that information, the extent to which one might agree with the proposition that someone's point is valid will vary. For example, one person may regard a point as valid because they agree with it; another person may regard a point as valid because it is factually demonstrable; yet another person may regard a point as valid because it is clear and easy to understand. In each case, the assessment of validity is based on a different set of criteria: opinion, fact-checking, or communicative effectiveness. In other words, we cannot make an objective judgment that a test is valid or invalid; rather, we can only make judgments that a given test is more or less valid for which specific purpose, of which version of a construct, or toward what kinds of effects. For this reason, there is no algorithm or criterion or methodology that can serve as a rubric for assessing validity. Rather, judgments of validity are inferences; validity is judged on the basis of inferences about purposes, constructs, and beliefs about what counts as operationalization of any given concept.

Regardless of one's agreement with such a point, *valid* carries with it connotations of power, as it tends to codify a particular thing as sound, as fact, or as knowledge. In the Foucauldian (1980) sense, it signifies to us that something is True (note the capital *T*). Although some scholars dismiss the pursuit of validity in scientific research (Wolcott 1990; Lather 1993; Gergen and Gergen 2000), the characteristic is widely used in the field of educational measurement, where validity refers to the alignment between what a test measures and what it claims to measure.

With roots among psychologists studying intelligence and cognition more broadly (e.g., Terman et al. 1915; Thorndike 1916), the meaning of validity and the process of assessment validation has evolved significantly over the past century, from purely statistical validations of assessments (e.g., factorial validity) to checks of differing types of validity (e.g., content validity, predictive validity), among other approaches (Sireci and Sukin 2013). Today, though there is still debate (Newton and Baird 2016), validity largely centers on how well a test measures what it claims to measure (Kane 2012; Newton 2012; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). That is, rather than breaking validity into constituent parts, validity is a unitary concept that "refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of a test" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014, 11). In this way, validity is not broken into a binary of valid/invalid because, regardless of the construct of interest, once we move from construct definition to its operationalization in an assessment, perfection is not feasible. Validity of a given assessment, then, falls along a spectrum of persuasion.

Authors of the *Standards for Educational and Psychological Testing* (a book hereafter referred to as the *Standards*) synthesize perspectives on what counts as persuasion and provide guidance for individuals seeking to validate an assessment (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). Per the *Standards*, there are five categories of evidence one might draw from (i.e., infer) for validation. These categories address:

1. Assessment content (the extent to which an assessment aligns with the construct of interest),
2. Response processes (test takers should engage with the assessment in ways test developers and the construct anticipate),
3. Internal structure (if some aspects of the construct are to be distinguished, or if the test is to function differently for different groups, there should be evidence for these patterns),
4. Relations to other variables (if the construct of interest relates to external variables, or if construct performance is to generalize to other contexts, evidence should support those propositions)
5. Consequences of the assessment (benefits of the assessment should outweigh its consequences).

It is jointly incumbent on the test maker and test user to provide combinations of these sources of evidence when validating their assessment.[3] The authors of the *Standards* establish this imperative early on, stating that "Evidence of the validity of a given interpretation for a specified use is a necessary condition for the justifiable use of the test" (11). Similarly, Kane (2012) notes: "If a lot is being claimed, a heavy 'burden of proof' is imposed on those making the claims" (70). That being said, there is no combination of these five sources that produces a valid assessment. The validation process varies based on inferences about the assessment itself, the meaning assigned to its outcomes, and the potential use of such outcomes. For example, the validation process of a university's mathematics placement exam will be different if exam score interpretations are taken as suggestions versus if they rigidly influence a student's course options; the validation of the same exam will be different yet if the construct of interest is quantitative literacy versus mathematical literacy. We do not talk about the validity of the assessment itself, but rather the validity of the assessment within the broader milieu in which it is administered.

To summarize, then, the validation process for an assessment is contingent upon a variety of factors, including what the test purports to measure, how scores are interpreted, and what the consequences are of such interpretations. The five evidence sources discussed above collectively contribute to the justification of proposed interpretations for proposed uses. Later, I will revisit the five evidence sources above in discussing my external validation of the PIAAC numeracy

[3] Note the developer and user may be the same individual or collective.

assessment. Note that I use the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014) as the guiding framework for validation, rather than derivative frameworks like Evidence-Centered Design (Mislevy and Haertel 2006) that specify a means of validation, as the *Standards* are broader in scope.

Assessing Numeracy

In the context of quantitative literacy or related constructs, assessment is not a novel concern (Cahoon and Kiliç-Bahi 2019). As a new skill for the twenty-first century, or a new requirement in postsecondary general education programs, quantitative literacy is a construct that administrators, faculty, and policymakers at multiple levels express increasing interest in surveilling. For example, we see this interest manifest in

- the creation of several VALUE rubrics from the Association of American Colleges and Universities, one of which centers on numeracy (Rhodes 2010);
- the recent creation of the HEIghten® assessment of quantitative literacy for postsecondary institutions from the ETS (Roohr et al. 2017);
- a National Science Foundation grant awarded to multiple institutions for the development of a numeracy assessment instrument (Gaze et al. 2014);
- the numeracy assessment on PIAAC, and even a special year of PISA devoted to numeracy (Kosko and Wilkins 2011; Gal and Tout 2014);
- the inclusion of a numeracy domain in the Collegiate Learning Assessment (CLA) (Klein et al. 2007); and
- a special issue on assessment in *Numeracy* (Vacher 2015).

The projects and scholarship listed above represent only a sample of efforts to assess numeracy; they vary in goal, format, funding (or lack thereof), and conceptualization of numeracy, among other things. Regardless of the flourish associated with these assessments—including multi-million-dollar funding, white papers, external publications, and an uptake in media sources—the *Standards* suggests that results from these assessments have little substantive meaning without accompanying discussions of validity (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). As I will argue below, assessments of numeracy (operationalized through a competency perspective) are especially tenuous, as the setting and assessment itself fundamentally obfuscate the construct of interest. An implication of this proposition is that—as numeracy researchers and scholars—we should be particularly demanding in thinking through the validation process of assessments we develop.

Challenges to Numeracy Assessment

Assessments of any construct are necessarily only proxies for that construct, unless those assessments are practical, real-life, real-time engagements. Scholars developing written assessments involving the construct of numeracy face a special hurdle to the first source of evidence in the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014), in that the construct nearly always addresses some notion of the real world that is somehow separate, spatially or temporally, from the writing of the definition of numeracy. Similar issues arise in the assessment of constructs such as critical thinking (Rear 2019) or problem-solving (Griffin et al. 2018). In contrast, the assessment of skills, such as the ability to graph a rational function or describe the steps of meiosis, is less tenuous, as no claim is made about when and how these skills might manifest. This is not to imply that the development of a numeracy assessment is impossible, because as noted earlier, validity is not only about construct validity. However, insofar as assessment content feeds into the development of interpretations and proposed uses of test scores, claims of validity require that the content align with interpretations that use language about that construct.

Synthesizing the diverse ways scholars have used terms like numeracy, quantitative literacy, and quantitative reasoning, Karaali et al. (2016) converged on a common "thread," stating that the terms tend to connote "a competence in interacting with myriad mathematical and statistical representations of the real world, in the contexts of daily life, work situations, and the civic life" (25). As one might imagine, the inherent grounding of the three terms in the "real" differentiates them from other things one might assess, such as the ability to factor a polynomial, where the assessment setting and construct setting (though ambiguous or not provided at all) are likely to align more closely. A host of scholars (e.g., Grawe 2011; Kosko and Wilkins 2011) have discussed this distinction at length, arguing in essence that numeracy assessments with limited response options (e.g., multiple-choice questions, numerical entry questions) fail to capture the essence of the real in numeracy. These scholars suggest that other mediums, such as essays or portfolios (Klein et al. 2007; Grawe et al. 2010; Rhodes 2010; Shavelson et al. 2019; Zerr 2019), are better suited for capturing what one means by numeracy. Though the aforementioned scholars do not take on a social practices perspective of numeracy explicitly, the issue they tackle—that of capturing the real—is explained well through such a perspective.

To expand on this point, we can interrogate the notion of *competence* included in Karaali et al.'s (2016) statement. The inclusion of competence in their thread suggests a functional or skills-based approach to the terms, meaning that, when evidenced through action, numeracy, quantitative literacy, and quantitative reasoning, all hinge in some way on some subset of skills (e.g., the ability to convert

from a decimal to a percentage). If the construct we seek to understand is what it is that people actually *do* with numbers, the definition itself of that action should not hinge on ability. Drawing from scholars largely in the anthropology and literacy studies communities, Oughton (2018), and later Craig and Guzmán (2018), challenged a functional view of numeracy in favor (or acknowledgement) of a practices-oriented view. A practices approach to numeracy views numeracy through the lenses of *practices* and *events*. Craig and Guzmán define numeracy events as events which are mediated in some way by quantification; such events are observable insofar as they "happen," whether mentally or physically (2018). From this definition, numeracy practices are those patterned (or repeated) things individuals tend to do in numeracy events, coupled with the significance individuals ascribe to such events. Distinct from a functional approach to numeracy, where numeracy is viewed as a set of skills used in context, "A social practice perspective not only takes into account different practical contexts; it also considers how people's life-histories, goals, values and attitudes will influence the way they carry out numeracy" (Oughton 2018, 6). Oughton's remarks are corroborated by a variety of studies in the context of numeracy (Carraher et al. 1985; Lave and Wenger 1991; Kahan et al. 2017; Tunstall et al. 2018) that suggest that skills alone do not dictate the nature of numeracy events.

Indeed, a central benefit of this perspective is that it acknowledges that our actions in the world outside of formal assessments are complex and ill-defined. Moreover, it disputes any assumption that ability (as measured by a test score) determines action, given that actions are influenced by more than just ability. Hence, if an assessment of numeracy only addresses ability, it raises fundamental questions of validity, that is, whether the test measures what it claims to measure. Due to a dearth of resources or a desire for efficiency, groups or researchers may be forced to resort to assessments that may be quickly administered and scored (PIAAC Numeracy Expert Group 2009; Shavelson et al. 2019), but the analysis here will contribute to conversations about the validity of such an assessment with respect to the interpretations and uses of the assessment. In short, especially when testing policies prioritize expediency, they have the potential to marginalize issues of validity in the process. In the analysis that follows, I adopt a social practices view while recognizing that I cannot change the construct that PIAAC developers intended to measure in the numeracy portion of the assessment. This framework for numeracy will manifest when I discuss or assess claims that link scores with action, as well as when I use the term *practices* or *events* to describe particular PIAAC components.

Organizing This Exploration

In following the path set out by the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014), questions that formally guided this investigation were: (1) What does the PIAAC numeracy assessment claim to measure? (2) What are the intended uses of the assessment? (3) How are we to interpret scores with those uses in mind? And (4) to what degree do evidence and theory support interpretations for those uses? Though the first three questions require research, the fourth question is the central research question of this investigation (and invites analysis more so than summary). Taken together, answers to these four questions allow me to talk about the validity of PIAAC numeracy assessment scores with respect to their intended use. Because this analysis is intended to provide readers with insights into their own assessment practices and development, I encourage the reader to consistently reflect on how this work would apply to other contexts outside of the PIAAC.

Method

Data Sources. In addition to several analyses of results, the OECD provides various resources for those interested in understanding how the PIAAC numeracy assessment was conceptualized, designed, and then implemented. These sources are available from the OECD's iLibrary, which hosts thousands of books, working papers, policy documents, and data sets, and serves as "the gateway to OECD's analysis and data."[4] To find documents reporting the PIAAC numeracy assessment, I used the iLibrary's search engine and the terms PIAAC and numeracy, compiling all documents that reported on the conceptualization, design, or implementation of the numeracy assessment. The initial search using the terms *PIAAC* and *numeracy* yielded 1,092 results, many of which were not related to what I was searching for, so it was necessary to delimit the search to documents (not datasets alone, for example) written in English (some documents in the database are written in French), and then to cull from those results documents that concerned the conceptualization, design, or implementation of the numeracy assessment. If a document referenced a previous OECD-published document to describe any of those elements, I did not include the newer document in the documents that I analyzed. This search process ultimately yielded

- a report from the PIAAC's Numeracy Expert Group (2009),
- an overarching framework document describing the constructs of interest in PIAAC (OECD 2012),
- a comprehensive "Technical Report" describing the minutiae of the development process (OECD 2016a),

[4] See https://www.oecd-ilibrary.org/.

- a *Reader's Companion* to the PIAAC's development (OECD 2016b), and
- a detailed *First Results* document from the 2011–2012 administration of the exam (OECD 2013a).

The number of pages in each of these documents, by order of bullet points, was 67; 62; 1,233; 130; and 466. Specifically unavailable to the public, though, are the 56 items used in the Numeracy Assessment. The OECD data request team did not grant me private access to the items (despite stating that I would not share them with others). Five of the fifty-six items (reportedly representative of the larger set) are available to the public through an informal document[5] on the PIAAC site and a simulation[6] of the actual assessment. Finally, it is worthwhile to point out that I only included documents pertaining to the first administration of the PIAAC, and not any documents pertaining to the upcoming second administration. Where relevant in the analysis, I still describe newer literature—both from members of the original Numeracy Expert Group (Tout et al. 2017) and other scholars (Evans 2014)—that highlights any strengths or limitations of the original assessment.

Analytical Framework. I answer the first three research questions using data from the sources bulleted above. The means by which I analyzed data to answer those questions are discussed in their respective sections below. The fourth research question—that of the extent to which theory and evidence support interpretations with respect to the proposed assessment uses—invites an evaluative argument based on sources both internal and external to the OECD's iLibrary. For this last analysis, I drew from relevant sources of validity evidence, as described in five broad categories of the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). I repeat those evidence sources below, this time parenthetically including commentary specific to the PIAAC numeracy assessment:

1. Assessment content (the numeracy construct description should align with its operationalization via test items; though there are only five publicly available items, these are reported as being representative of the larger set),
2. Response processes (if test developers expect test takers to engage in numeracy in specific ways, evidence should support that questions elicit that behavior),
3. Internal structure (for example, if assessment items are to be of increasing difficulty, evidence should support that assumption),
4. Relations to other variables (if other variables, such as literacy assessment score, are known to relate to numeracy, then evidence should support that the numeracy assessment differentiates those constructs), and
5. Consequences of the assessment (if there are to be material consequences of an individual or country's score on the numeracy assessment, then evidence should support that those consequences follow from differential scores on the assessment).

[5] See http://www.oecd.org/skills/piaac/Numeracy%20Sample%20Items.pdf for the sample items.
[6] The simulation is available at http://www.oecd.org/skills/ESonline-assessment/.

As noted earlier, not all five categories may be relevant—the evidence needed will depend on answers to the first three questions.

In following the *Standards*, for assessment content I consider construct validity, i.e., the alignment between the construct and example assessment items (noting the limitation of the analysis); doing this entails examining available assessment items to compare what is assessed to what is intended to be assessed in the construct. For response processes, I discuss whether evidence—such as field testing or pilot studies—is presented by test developers to suggest that test-takers indeed engage in processes expected of numerate behavior. For relations to other variables, I looked within the five key PIAAC documents to see if theoretically related variables such as literacy and mathematical skills (variables which I chose, as explained below) are considered by PIAAC developers in relation to the numeracy construct. As noted by the authors of the *Standards*, it is important that evidence be provided that demonstrates that the assessment of a construct *X* theoretically related to another construct *Y* is indeed measuring *X* and not *Y*. Finally, for consequences of score interpretations, I discuss whether evidence is provided by test developers in the five PIAAC documents to justify that score differentials correspond to actions based on interpretations of those scores. Note that nearly all of these sources of evidence require that I look for their presence in documentation literature concerning the PIAAC. In the relevant parts of the section that follows, I describe how I looked for this specific evidence within PIAAC documentation. Taken together, consideration of these five categories provides evidence of the extent to which we might be persuaded that the score interpretations from the PIAAC assessment are justified in light of the test's proposed uses.

Analysis

I organize the analysis in relation to the four research questions in two parts: those related to questions one through three, and those related to question four.

Interpreting a Measurement for a Specified Use

What the PIAAC Numeracy Assessment Measures. To answer the first question, that which the PIAAC numeracy assessment attempts to measure, I began by examining an OECD white paper from its PIAAC Numeracy Expert Group (2009) for descriptions of what the numeracy portion of the PIAAC attempts to measure. I used this document as the primary source of evidence for answering this question, given that it is the sole OECD document delineating the numeracy construct and is referred to by testmakers in other documents when describing the numeracy portion of the PIAAC. Given the document's organizational structure (described in further detail below), answering this question entailed summarizing the authors' argument rather than looking through the document for specific codes (for example) related

to what the assessment might measure. I referred to other documents, including the Technical Report (OECD 2016a), which describes in detail the test development process, and the *Reader's Companion* (OECD 2016b), which outlines the test for those interested in its results, for conflicting information concerning what the numeracy assessment measures. For example, it could have been the case that the test developers decided to include only certain parts of the numeracy construct as outlined by the Numeracy Expert Group. In that sense, conflicting information could manifest as explicit statements suggesting that the construct assessed was distinct from that which the Expert Group described. There were no major deviations in the design or enactment of the first administration of the PIAAC reported following the Expert Group's (2009) publication.

In the 67-page document, the group situated their conceptualization of the construct of numeracy within those from other groups, assessments, and constructs (e.g., mathematical literacy). Ultimately, the group arrived at a two-pronged definition, with the first prong being that "Numeracy is the ability to access, use, interpret, and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life" (21). The authors noted their intentionality in using the word *engage* in the definition, stating that numeracy necessarily involves dispositional elements beyond just skills. To the authors, these dispositional elements include "positive beliefs and attitudes about mathematics and about oneself as a person capable to cope with mathematical tasks" (24).

Going further, the authors stated that because numeracy is a complex construct, it was essential to add to the definition of the notion of *numerate behavior*. Numerate behavior "involves managing a situation or solving a problem in a real context, by responding to mathematical content/information/ideas represented in multiple ways" (21). According to the authors, this expansion of the definition allowed for actual operationalization in an assessment, "thereby contributing to the assessment's validity and interpretability" (21). That is, the expanded definition was an important contributor to the assessment's validity. Despite this claim, the authors did not discuss validity elsewhere in the document. With that said, the authors did discuss how the introduction of the phrase *numerate behavior* contributed to the assessment's operationalization. The definition of numerate behavior was then operationalized through questions that drew from

- four categories of *real contexts* (e.g., everyday life, work)
- five types of *responses* (e.g., interpret, communicate)
- four domains of *mathematical content/information/ideas* (e.g., dimension and shape), and
- six venues for *multiple representations* (e.g., maps, tables) which would guide the development of their assessment items.

Importantly, the numerate behavior outlined above hinges on the "activation of" "enabling processes," which include

- mathematical knowledge and conceptual understanding
- adaptive reasoning and mathematical problem-solving skills
- literacy skills
- beliefs & attitudes
- numeracy-related practices and experience[, and]
- context/world knowledge (22).

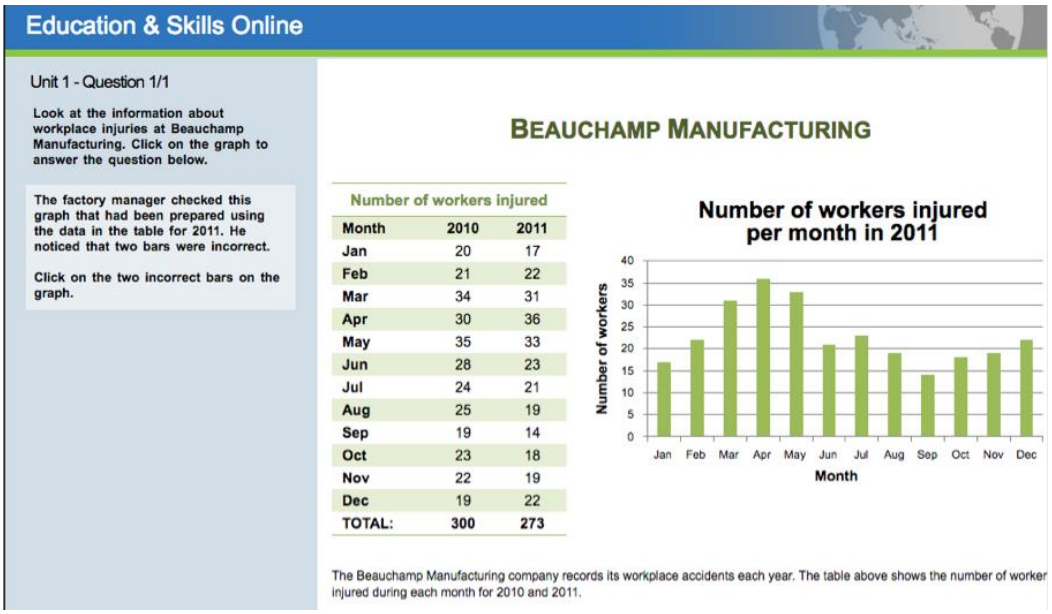
Where relevant, I expand on the ideas in the two bulleted lists above. The enabling processes will be particularly important for discussing interpretations of scores. For now, I have discussed how the Expert Group used its two-pronged definition to attempt to operationalize numeracy through the notion of numerate behavior.

With that definition in hand, the document then describes how such a framework might manifest through the actual assessment. To that end, it includes a discussion of the limitations of the PIAAC testing environment and how that environment influenced the creation of their assessment item pool. In particular, the eighty-minute test (including all questions, as well as background surveys) was to be given at home, with a proctor present, using a computer and automated scoring. Those constraints led the Expert Group to create an item pool where principles guiding item creation were that the items cover as many mathematical domains as possible, have "maximal authenticity and cultural appropriateness" (which is a validity claim), be scored automatically, cover different levels of difficulty, require different response actions (e.g., interpret versus compute), be time efficient (i.e., answerable quickly), and adaptable without significant modifications across participating countries (36–37). In my view, the Expert Group faced a tall task, and I discuss the extent to which they (in my view) successfully worked within and around such constraints in the context of validation later in this paper. An example of an assessment item is provided in Figure 1 below. Other publicly available items are provided in the Appendix.

The "Beauchamp Manufacturing" problem requires the test taker to identify two bars on a bar graph that are apparently incorrect in light of the table the data is based on (as opposed, for example, to identifying places where data in the table itself might be incorrect). In relation to the Expert Group's framework for numerate behavior, note that the context here is work; the response type is to interpret and evaluate, as the respondent must interpret the bar graph and then evaluate aspects of its accuracy; the item falls under the grouped mathematical domain of data and chance; and the representation includes both a table and bar graph. The sample item demonstrates the goals the Expert Group discussed in creating problems, as it is quickly answerable, automatically graded, grounded in a potentially authentic

context, and adaptable across countries (e.g., bar graphs do not vary significantly in other countries).

The constraints that the Expert Group acknowledge, and that we see manifest in the item in Figure 1, invite critique concerning the apparent disconnect between numeracy as a complex construct—a behavior contingent on enabling processes like beliefs and attitudes—and one that could somehow be operationalized in the manner described above. The authors recognize this issue and include disclaimers throughout their writing. For example, after discussing the constraints above, the Expert Group notes: "As a result of the restrictions discussed above, certain types of numeracy tasks, especially those involving interpretation or evaluation/analysis with communication responses, receive only partial or slight coverage in the first cycle of PIAAC" (34). As I discuss in the next section, the extent to which this complexity and hedging manifests in other aspects of the test development, such as interpretations of or uses of scores will vary. In summary, to the question of validity, that is what the PIAAC numeracy assessment aims to measure, the answer—subject to hedging—is numerate behavior, which the Expert Group categorizes as falling along dimensions of context, response type, mathematical content, and representation medium.



Month	2010	2011
Jan	20	17
Feb	21	22
Mar	34	31
Apr	30	36
May	35	33
Jun	28	23
Jul	24	21
Aug	25	19
Sep	19	14
Oct	23	18
Nov	22	19
Dec	19	22
TOTAL:	300	273

Figure 1. Publicly Available Numeracy Item from the PIAAC. Source: http://www.oecd.org/skills/ESonline-assessment/takethetest/?.

Uses of the PIAAC Numeracy Assessment. To search for purpose, or the intended uses of the assessment, I examined the five key documents that the previous search process had yielded. In examining those documents, I looked for signaling words such as "purpose" or "objective" and an explicit declaration of that purpose or objective in the context of all of the PIAAC (e.g., not just the literacy portion). Because not all declarations of purpose contained such signal words, though, it was important to read each document more than once for this specific search. For example, in the beginning chapter of *Literacy, Numeracy and Problem Solving in Technology-rich Environments: Framework for the OECD Survey of Adult Skills* (OECD 2012), "Why Assess the Skills of Adults?" the authors opened with the statement:

> Understanding the level and distribution of these skills among the adult population in participating countries, as well as the ways such skills are developed and maintained, and the social and economic benefits for individuals, is important for policy makers in a range of areas of social and economic policy (1).

The statement preceding "is important for" suggests what the OECD attempts to do through its assessment. Specifically, in this OECD document, judgments of validity are tied to "social and economic benefits for individuals." The primary document that proved fruitful from those five documents was *The Survey of Adult Skills*: *Reader's Companion* (OECD 2016b), which had the explicit motive of describing the "'what' and 'how'" of the PIAAC (13). In a manner similar to my approach in answering the first question, I later corroborated my findings by looking for confirmatory and dis-confirmatory evidence in the five sources. I did this by re-reading the five documents to look for statements that suggested a purpose or use either similar to or contradictory to those that I had initially found. Ultimately, I found the purposes bulleted below; these were stated as the major analytical objectives of all of PIAAC:

- Determine the level and the distribution of proficiency in key information-processing skills for certain subgroups of the adult population.
- Better understand factors associated with the acquisition, development, maintenance, and loss of proficiency over a lifetime.
- Better understand the relationship of proficiency in information-processing skills to economic and other social outcomes. (36)

These objectives are found somewhat less explicitly in other OECD documents (cf. OECD 2012, 1), but note that none of the documents I examined contained evidence suggesting that these were not the uses of the PIAAC.

The list above concerns objectives of all of the PIAAC (i.e., the assessments of literacy, numeracy, and problem-solving in technology-rich environments), and the references to *information-processing skills* suggests that one might read the list with the construct of numeracy explicitly in mind. Note that the PIAAC developers intended to meet the first objective through the three domain assessments, and the

second and third objectives through the domain assessments coupled with the background questionnaire, which included closed-response questions about the frequency and use of various skills in one's life, as well as closed-response questions about one's health, occupation status, and other elements related to economic and social outcomes. Beyond these direct uses of the assessment scores, the ultimate goal of PIAAC is to "identify levers" in order to "reduce deficiencies," the rationale being that "Skills transform lives, generate prosperity and promote social inclusion" (OECD 2013b, 4–6). While the notion of identifying levers relates to the bulleted objectives, the task of reducing deficiencies and the rationale for doing so are beyond the scope of what assessment scores can do alone.

Interpreting PIAAC Numeracy Scores. Through the third research question I ask one of the fundamental questions of validity for the PIAAC instrument: In light of the purposes outlined above, how is one to interpret scores on the numeracy assessment? Taken together, the technical report (OECD 2016a) and *Reader's Companion* (OECD 2016b) shed light on score interpretations. The administration of the PIAAC was a multilateral effort, with dozens of individuals from the ETS, OECD, and partner countries working together to develop and administer the exam. From a methodological standpoint, an important point to note is that—as stated in the first analytical objective above—test developers sought the distribution of skills proficiency among *subgroups* of the adult population—not to report (or even provide) results at the individual level.[7] Using Item Response Theory scaling and latent regression modeling, test developers created proficiency scales for each of the three domains of interest: literacy, numeracy, and problem-solving in technology-rich environments. Each of the scales ranged from 0 to 500 points, and every task in the numeracy domain fell at a point along that scale to indicate its difficulty based on field pilots of the assessment items (OECD 2016a). Test developers then combined item difficulty information with performance information on groups and subgroups within each country, the goal being to develop an "ability distribution" for relevant groups in specified domains (OECD 2016a, 579). To facilitate interpretation of the distributions, each 0–500 scale was broken into six levels: Below Level 1, Level 1, Level 2, and so on until Level 5. Because these proficiency levels are central to how scores are reported, I include those for the numeracy assessment in Table 1 below.

[7] Individuals did not receive score reports, nor counseling or other resources for improving the skills tested.

Table 1
PIAAC Numeracy Proficiency Levels

Proficiency Level	Description
Below Level 1 (0 to 175)	Tasks at this level are set in concrete, familiar contexts where the mathematical content is explicit with little or no text or distractors and that require only simple processes such as counting, sorting, performing basic arithmetic operations with whole numbers or money, or recognizing common spatial representations.
Level 1 (176 to 225)	Tasks in this level require the respondent to carry out basic mathematical processes in common, concrete contexts where the mathematical content is explicit with little text and minimal distractors. Tasks usually require simple one-step or two-step processes involving, for example, performing basic arithmetic operations; understanding simple percents such as 50%; or locating, identifying and using elements of simple or common graphical or spatial representations.
Level 2 (226 to 275)	Tasks in this level require the respondent to identify and act upon mathematical information and ideas embedded in a range of common contexts where the mathematical content is fairly explicit or visual with relatively few distractors. Tasks tend to require the application of two or more steps or processes involving, for example, calculation with whole numbers and common decimals, percents and fractions; simple measurement and spatial representation; estimation; and interpretation of relatively simple data and statistics in texts, tables and graphs.
Level 3 (276 to 325)	Tasks in this level require the respondent to understand mathematical information which may be less explicit, embedded in contexts that are not always familiar, and represented in more complex ways. Tasks require several steps and may involve the choice of problem-solving strategies and relevant processes. Tasks tend to require the application of, for example, number sense and spatial sense; recognizing and working with mathematical relationships, patterns, and proportions expressed in verbal or numerical form; and interpretation and basic analysis of data and statistics in texts, tables and graphs.
Level 4 (326 to 375)	Tasks in this level require the respondent to understand a broad range of mathematical information that may be complex, abstract or embedded in unfamiliar contexts. These tasks involve undertaking multiple steps and choosing relevant problem-solving strategies and processes. Tasks tend to require analysis and more complex reasoning about, for example, quantities and data; statistics and chance; spatial relationships; change; proportions; and formulas. Tasks in this level may also require comprehending arguments or communicating well-reasoned explanations for answers or choices.
Level 5 (376 to 500)	Tasks in this level require the respondent to understand complex representations and abstract and formal mathematical and statistical ideas, possibly embedded in complex texts. Respondents may have to integrate multiple types of mathematical information where considerable translation or interpretation is required; draw inferences; develop or work with mathematical arguments or models; and justify, evaluate and critically reflect upon solutions or choices.

Source: Proficiency descriptions in this table are taken directly from OECD (2016a, 588–591).

Test developers arrived at these proficiency scales for each of the three domains using standard test-norming procedures: upon aggregating performance data and meeting with the domain expert groups to discuss characteristics of the assessment items. Though individuals did not receive their own scores, the developers state that the score of an individual falling at a particular proficiency level (e.g., Level 4, and in particular, the score 330) indicates that the person would be expected to correctly answer task items with a difficulty level of 330 about 67% of the time.[8] The "Beauchamp Manufacturing" problem from Figure 1 falls into Level 2 from those levels given in Table 1, as it has few distractors (i.e., one column

[8] This quantity, 67%, is referred to as a response probability (RP) value.

of data is irrelevant), requires only estimation, and does not involve several steps. To provide an example of an interpretation of these scores, I draw from a "Summary of Findings and Policy Recommendations" from *Time for the U.S. to Reskill? What the Survey of Adult Skills Says* (OECD 2013c). The first key finding leading off the document is the following: "Low 'basic' skills (literacy and numeracy) are more common in the United States than on average across countries" (11). The statement itself relates to the first purpose of the PIAAC outlined in the three objectives earlier—that of determining "the level and the distribution of proficiency in key information-processing skills for certain subgroups of the adult population" (OECD 2016a, 36). The interpretation of this statement is that the percentage of the US adult-aged population scoring at or below Level 1 on the numeracy scale is greater than that of the average across other countries tested. Similar statements can be said about literacy levels.

In answering questions one through three, I have discussed the construct of numeracy that the PIAAC's developers sought to measure, the stated uses of the numeracy assessment, and the interpretations one is to make based on scores on the numeracy assessment. In an argument-based approach to validation, the core of the validation process is to then consider the extent to which interpretations for those uses are justified in the context of what developers seek to measure. Thus, in the next section, I take this information to answer my research question: To what extent do theory and evidence support interpretations for those uses?

Supporting Interpretations with Theory and Evidence

The first three questions invited summary more than analysis or evaluation. In considering how evidence and theory support interpretations for specified uses, the task transitions to one of making or evaluating claims about support for those interpretations. As one might imagine, the universe of possible interpretations of scores with respect to the three overarching objectives of the PIAAC is vast. Given the reams of work produced by the OECD in describing the PIAAC and its development, any consideration of validity would necessarily be vast as well. I restrict my scope here to interpretations of the PIAAC numeracy assessment scores as they relate to objective one of the PIAAC (determine the level and the distribution of proficiency in key information-processing skills for certain subgroups of the adult population). The rationale for that specific restriction is that objective one centers around the numeracy assessment itself, whereas objectives two and three focus on its relation to the background questionnaire—a component of PIAAC that, while potentially interesting to study, is not the numeracy assessment itself. In my closing discussion, I will revisit possibilities for future work in relation to opening up the validity discussion to those involving objectives two and three.

I structure this section into parts corresponding to sources of validity evidence discussed in the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). As I already noted, not all assessments invite the same types of validity evidence, so some sections will be shorter than others. For example, the category of *internal structure* in this context is not fruitful to explore, because the PIAAC numeracy portion does not include composite or subtest scores to measure different aspects of the numeracy construct. The PIAAC Numeracy Expert Group (2009) made no claims that the numeracy assessment measures multiple constructs; the only claim made relative to internal structure is that some items were more difficult than others, based on a collection of factors related to item complexity. Such claims were substantiated through pilot evidence and discussions among members of the Expert Group (OECD 2016a), so I do not devote space here to that source of evidence. Rather, I focus here on the categories of assessment content, response processes, relations to other variables, and consequences of score interpretations.

Assessment Content and Response Processes. The PIAAC Numeracy Expert Group (2009) described in detail their conceptualization of numeracy as it should manifest in the assessment item pool. With respect to the operationalization of the construct—that is, the assessment items themselves—the group used the notion of *numerate behavior* to facilitate item development. As noted earlier, numerate behavior "involves managing a situation or solving a problem in a real context, by responding to mathematical content/information/ideas represented in multiple ways" (21). Built into the expanded version of this definition are response processes (e.g., interpret, communicate), so I group that category of validity evidence into this discussion as well. The item provided in Figure 1, the "Beauchamp Manufacturing" problem, is an exemplar of the construct of numerate behavior operationalized in an assessment task. Accompanying each of the five publicly available items is a similar mapping from the definition of numerate behavior to an actual task. In combining the developers' discussion of numerate behavior with the tasks publicly available and the statistical techniques used to determine scores, there are no salient concerns, writ large.

That being said, in light of the test maker's first objective of determining the level and distribution of numeracy within and across populations, the primary concern that arises in considering the content of the assessment is in *how* the test items purportedly align with the instrument's stated definition of numerate behavior. In particular, I argue below that the test items do not account for what it could mean to engage in numerate behavior as delineated by the Expert Group—an argument that, since the first administration of the PIAAC, has been developed in a similar way by several original members of the Expert Group (see Tout et al. 2017). This inability to account for the possibilities of numerate behavior goes

beyond what one might expect of any assessment by virtue of its nature as a proxy. To justify this claim, note that there are three key phrases within the definition of numerate behavior that invite critique here: "managing a situation or solving a problem," "real context," and "by responding to." Below, I expand on how, upon further inspection, these aspects of the construct are not adequately captured in the assessment items.

With respect to "managing a situation or solving a problem," it is essential to note that judgments about management are inherently bound to a context. Through a social practices lens of numeracy, one would say that for the test taker, the context of these problems is the context of being on a computer and answering questions while being observed by an interviewer (as is the case with any similarly-structured assessment). It is not the case that the test-taker is actually at work and looking for errors in their bar graph. That is, the numeracy event occurs in answering the question, not in actually being in the world described in the question. Consider the question in Figure 1, which also appears in the first row of Table 2—looking at a bar graph for errors in one's work (or in this case, someone else's).

Table 2
Additional Context Considerations for Sample PIAAC Numeracy Items

Test Item	Description of Problem	Real-life Factors or Questions to Consider
Beauchamp manufacturing	The test-taker is asked to compare a bar graph with a table that generated that bar graph; the task is to determine which bars on the graph are incorrect.	If the bar graph is generated automatically from the table, is it realistic that only two bars would be incorrect? Would a person in this situation have coworkers that might be interacting with the presentation and that might be responsible for noticing the error as well?
Running shoes	The test-taker is provided with prices for two pairs of shoes, and asked to calculate the cost of the purchase if there is a discount for purchasing both pairs.	When making a purchase online, prices are often automatically calculated in the person's shopping cart. Does successfully managing a shoe purchase require knowing how to calculate this cost? How might a person's goals for the total purchase make this question more complex?
Temperature dial	The test-taker is presented with a temperature dial, and asked what the temperature would be if it were actually 30 fewer degrees Celsius.	Because many temperature gauges are now digital, how might this problem be different? In what context would someone be reading a dial that is incorrect by 30 degrees Celsius, and is it the case that the problem in that context would be knowing what the new temperature would be?

Source: These items are available in the Appendix (see http://www.oecd.org/skills/piaac/Numeracy%20Sample%20Items.pdf).

The way that one responds to such a "problem" is mediated by a variety of factors, notably including what is expected of them (in being positioned as a test-taker, the expectation is that they will answer questions "correctly"). There is no room provided for the test-taker to respond to the situation, to ask questions, or to situate their own views, knowledge of the context, beliefs, or habits in relation to the task. They are to simply find two incorrect bars on a graph. In Table 2 above, I

raise similar points for two other publicly available questions. These questions are given to test-takers despite the fact that the PIAAC Numeracy Expert Group (2009), as noted earlier, specifically defined numerate behavior as being contingent upon certain enabling processes, which include beliefs, attitudes, as well as numeracy-related practices and experience (22). Given that extant research suggests that the ways one might attend to this situation would inevitably differ if encountered outside of this setting (Carraher et al. 1985; Lave and Wenger 1991; Kahan et al. 2017; Tunstall et al. 2018), what is it that we actually learn from seeing what one can do in this restricted context? I offer one potential answer to this question below, but do not fully answer this question in this paper.

It is assumed that one would respond (i.e., the definition of numerate behavior states "by responding to") by examining the bar graph in comparison to the table to find the error. However, in a context in which this problem actually arose outside of a test-taking setting, one might wonder if the expected mathematics (e.g., examining the bar graph) would be used at all (Oughton 2009). Given that the graphs were clearly generated by the use of a computer, I question how a computer would make such a mistake if it was relying on inputs from a table; of course, errors can occur, but their possibility does not make this sufficiently authentic in my view.

Beyond "managing a situation or solving a problem" and "by responding," the aforementioned remark speaks to the issue of "real context." Each of the problems on the PIAAC numeracy assessment is meant to emulate some real context. Through a social practices lens of numeracy, these contexts are real only insofar as they are real in the moment to the test-taker. Each task serves as a numeracy event. The extent to which that event occurs with some regularity outside of the PIAAC assessment—that is, for it to be a numeracy practice of the test-taker—is not clear, as evidence is not provided by the test developers. While most assessments deal in some way with the issue of the assessment being only a proxy for what one might do outside of the assessment environment, it is important to reiterate (as noted earlier) that there is a special hurdle for any assessment of numeracy to surpass given its inherent tethering to the "real."

The issue of "real" here may seem to be one of mere semantics, but it is essential to keep in mind that everyone's lived experiences, which ultimately are what numeracy practices in part capture, are different. Of course, it is possible that the assessment measures certain aspects or components of numerate behavior, but devoid of a fuller context and room for possibility in which that behavior might manifest, one is left to wonder (without any actual evidence) what only partial measurements tell us. It would be misleading then to claim that the assessment measures numerate behavior when the notion of what is real has not been properly qualified. Furthermore, though culture inevitably influences what is real to each of us, the test developers made clear that they sought contexts that supposedly apply to all cultures, stating: "Item content and questions should appear purposeful to

respondents across cultures, although it must be acknowledged that in a large-scale assessment such as PIAAC, not all items and contexts can be personally familiar to all adults within any one country, let alone across all countries" (PIAAC Numeracy Expert Group 2009, 35–36). In the context of what the assessment is supposed to measure, numerate behavior, it is essential to qualify how such statements influence what test scores actually mean (Evans 2014). Scores do not measure or tell us what the people in the representative population are doing, or what they might do in a situation, but instead, they tell us how well individuals might respond to a given artificial context to answer a question in a way that has been forced upon them. It does not tell us about the rich possibilities for nuance in response to situations that actually matter to adults. Again, these remarks then raise the question: what does the PIAAC numeracy assessment actually tell us about what people might actually do outside of the assessment setting? I cannot answer this question in the course of this analysis (alas, that is not the purpose of this paper), but I do discuss this issue in further detail in the Discussion.

Relations to Other Variables. A salient issue that one might anticipate in attempting to measure numeracy is in distinguishing it from other constructs. In the context of the PIAAC numeracy assessment, the definition of numerate behavior is that it involves using some type of mathematical information to manage a situation or solve a problem in a real context. In light of the discussion above, one might ask how the items used in PIAAC assess more than just the use of mathematical information to solve a problem. Put differently, one might ask, how are we sure that we are measuring numerate behavior and not just mathematical skills in isolation from numerate behavior more broadly? Furthermore, how do we know that the numeracy assessment is not a more elaborate assessment of literacy?

With respect to the former question—one that has been discussed in detail by scholars in quantitative literacy (see Steen et al. 2001)—the Numeracy Expert Group argues that contexts elevate these problems beyond that of context-free mathematics; however, they provide no empirical evidence (e.g., analysis to discern differences in responses to these question types) from the PIAAC or argumentative discussion to substantiate that claim. Calling attention to this absence is not meant to denigrate members of the Expert Group, but rather to point out that evidence necessary for validation is missing, and that we have room to grow if we are to develop assessments of constructs that hinge on relationships with other constructs. Indeed, across the five key documents that I examined in this study, I found no evidence (which would manifest as a statistical argument) that the numeracy assessment behaves differently than a more traditional mathematics assessment. The PIAAC Numeracy Expert Group (2009) explicitly acknowledges the latter question (from above), drawing from Baker and Street (1994) to suggest that the two constructs are not mutually exclusive. That being said, the Expert Group argues that numeracy "is a broad construct with a life of its own" and that its "skill levels

are not measured well by literacy measures" (8–9). Ultimately, the Expert Group's argument is that though numeracy tasks are embedded within texts, the tasks involve more than just reading, and that there are a host of enabling processes specific to numeracy, only one of which is literacy. With literacy, statistical evidence is provided related to the relationship between the numeracy and literacy assessments. Notwithstanding this argument from the Expert Group, the overall disattenuated correlation[9] in the initial round of the PIAAC from 2012 between countries' numeracy and literacy proficiency scores was 0.87 (OECD 2013a; OECD 2016a). Being above 0.85, this is a coefficient that some would suggest is sufficiently high to imply that the two measures are hardly discriminating different constructs (Clark and Watson 1995; Kline 2015). Despite this statistic, upon reporting these correlations, analysts noted, "Literacy and numeracy, nevertheless, constitute distinct skills, each defined by their respective frameworks" (OECD 2013a, 2). The statement inaccurately suggests that divergence in construct definitions is sufficient to establish divergence in construct operationalizations. I comment critically on this argument in further detail in the final section of this paper. In summary, of two important constructs that might co-vary with performance on the PIAAC numeracy assessment—mathematical skills more broadly, and literacy as operationalized on the PIAAC—we are not provided with sufficient evidence to support the notion that PIAAC numeracy assessment scores are valid for capturing numerate behavior.

Consequences of the Assessment. The last source of validity evidence discussed in the *Standards* includes consideration of consequences—direct and indirect—stemming from interpretations of scores for a given assessment (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). As discussed earlier, interpretations of PIAAC numeracy scores are meant to inform policymakers of the proficiencies of their constituents with respect to literacy, numeracy, and problem solving in technology-rich environments. Ultimately, a goal of PIAAC is to "identify levers" in order to "reduce deficiencies," the rationale being that "Skills transform lives, generate prosperity and promote social inclusion" (OECD 2013b, 4–6). Per the authors of the *Standards*, it is incumbent upon test makers to provide evidence that supports such logic (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014).

In the context of the chain of reasoning above, PIAAC developers would need to demonstrate that (a) interpretations of scores indeed provide evidence of deficiencies in the population of interest, and (b) once those deficiencies are addressed, nations and their "more proficient" constituents will be more prosperous

[9] Through disattenuation, one uses statistical information concerning reliability to correct for errors inherent in the measurement process (Osborne 2008).

and socially inclusive. The extent to which the developers demonstrated proposition (a) depends on how we hedge what is measured. As I have argued above, the PIAAC numeracy assessment has validity issues in its attempts to capture numerate behavior but may indeed have more validity for capturing numeracy skills in isolation of the broader enabling processes associated with those skills. With respect to (b), test developers rely on observational correlations between skills and income (among other metrics) that are based on a static dataset (i.e., the data are limited to one testing period).

If the developers are assuming a causal relation between improvements in PIAAC numeracy scores and metrics related to well-being—an assumption not directly stated, and that I cannot discern in the space of this analysis—then it is reasonable to suggest that they have not provided sufficient evidence toward that relationship. The assessment captures data on participants at one point of time, rather than longitudinally. Furthermore, the data are observational, rather than derived from any sort of controlled experiment. Existing research from scholarship on literacy suggests that a causal mechanism between literacy scores (on other assessments, not the PIAAC) and metrics related to well-being is misguided and not grounded in actual data (Graff 1978; Scribner and Cole 1981).

Finally, it is worth mentioning that validating discussions are typically found in reports of the assessment development process, and that evidence in relation to (a) and (b) are only in OECD score interpretation documents (OECD 2013a; OECD 2013c; OECD 2016b), rather than the development documents themselves (cf. OECD 2016a). Even where they do exist, the evidence in favor of (a) and (b) are never explicitly sectioned off (or even referred to) as validating discussions. This placement is not wholly surprising in the context of other developers' validations. In an analysis of assessments and associated validations from assessment developers, Cizek et al. (2008) found that this source of evidence was largely nonexistent in extant validations, despite the fact that key figures in scholarly discussions of assessment validation had called for its inclusion since 1989 (see Messick 1989).

Discussion and Looking Ahead

The end product of a validation process or study is not a yes or a no, but instead an inference based on a set of qualified statements about an assessment in the broader context of score interpretations for stated uses (Sireci and Sukin 2013; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). In this section, I synthesize my work above to make claims about the extent to which interpretations of scores on the PIAAC numeracy assessment are valid for the OECD's stated uses of the assessment. I then offer practical suggestions for those in the *Numeracy* community

interested in using or further exploring the PIAAC, or in developing their own assessments.

Beyond Valid or Invalid

Per the PIAAC Numeracy Expert Group, tasked with developing and operationalizing the construct of numeracy for the PIAAC, "Numeracy is the ability to access, use, interpret, and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life" (2009, 21). Going further, the group argued that such a definition is inadequate for conveying the construct's complexity and for operationalizing the construct through assessment items; for this reason, we need the notion of numerate behavior, which "involves managing a situation or solving a problem in a real context, by responding to mathematical content/information/ideas represented in multiple ways," and is contingent upon "activation of several enabling factors and processes" which include (among other things) beliefs, attitudes, practices, experiences, and real-world context knowledge (21–22). In each of the five publicly available numeracy items, test makers outline how the construct of numerate behavior manifests in the items.

In the discussion prior to this section, I outlined issues in how this operationalization manifests in an example assessment item, notably including that the assessment item itself (as representative of the others) does not allow for the enabling processes that numerate behavior is purportedly contingent upon. Furthermore, I critiqued the definition of numerate behavior itself, arguing that it assumes a binary notion of correctness in what it means for one to manage a situation or solve a problem (one that relies on mathematical behavior), and that it assumes a reality that only exists in the assessment itself. Though this critique suggests that the PIAAC assessment does not necessarily measure what it sets out to measure, and thus that assessment scores do not represent what was intended, it is important to keep in mind that validity is not just about construct-operationalization alignment, but rather about whether theory and evidence support interpretations of scores for proposed uses. In the context of the PIAAC numeracy assessment, a certain muddiness arises when we begin to consider how scores of the assessment are to be interpreted.

As noted earlier, numeracy scores are reported on the scale of proficiency given in Table 1. This scale was developed using pilot data and the Expert Group's comments on item difficulty. Based on this scale, scores about the construct of interest—numeracy, or numerate behavior—are ultimately then about the extent to which a group collectively answered a set of items varying in difficulty. Assuming that the experts involved in analysis completed their work correctly from a statistical standpoint (which I have no reason to doubt), scores, along with the interpretations provided in Table 1, appear to be valid for the use of describing the

skills discussed in those tables. The major caveat is that the numeracy suggested by the heading in the Table, and the construct purportedly measured and operationalized by the test developers, are different. Notwithstanding the potential validity of these specific score interpretations for a specified use, it is essential that one qualifies statements about the assessment itself so that individuals are not misled. If one examines the *Reader's Companion* (OECD 2016b), one sees in progression an overview of numeracy and numerate behavior, followed by the scoring table; there is no signaling that the two are in conflict. Hence, a potential consequence of score interpretations here is that one could be misled. For this reason, it is reasonable to argue that the validity of score interpretations is compromised.

In summary, the major finding pertaining to validity in this paper is that score interpretations from the PIAAC numeracy assessment may be considered valid for the use of describing distributions of proficiency in subgroups of interest, but

- the construct of interest—real-life numerate behavior—is not what is measured by the instrument,
- evidence distinguishing what is measured from other constructs, such as the OECD's conception of literacy, is largely absent, and
- consequences of the uses of the scores are not adequately justified.

These findings suggest some validity issues, namely that interpretations of scores do not align with descriptions of numerate behavior. Furthermore, they arise from my analysis of existing OECD documents and related literature—not from perusal of any straightforward discussion of validity from the test developers. The dearth of any validity argument from PIAAC test developers is a problem in itself, as it is incumbent upon test developers to clearly outline the evidence and theory that support interpretations of scores for specified uses (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014).

At this point, it is worthwhile to note that several members of the original Numeracy Expert Group have, in addition to noting issues in the original 2009 report, since worked to call attention to several shortcomings of the original PIAAC numeracy assessment administration to improve upon for the second administration currently in progress (Tout et al. 2017). While the group does an excellent job of describing issues, including (among other things) those related to the numeracy framework (e.g., calling attention to the need to account for a disposition to use numeracy) and assessment delivery (e.g., utilizing digital technologies), they do not directly connect their critique to its implications for the validity of the PIAAC numeracy assessment. Furthermore, in the few instances where they do discuss validity (e.g., Tout et al. 2017, 25), it is only done so in passing—without an explanation of what is meant by the term—and in the same sentence with reliability and fairness, a move that does not reflect the necessary foregrounding that validity

merits in the process of any assessment's development (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). While it is certainly laudable to make changes to an assessment to better capture the construct one intends to measure, there is a missed opportunity if these changes are not made in a manner that foregrounds validity.

Toward Caution and Responsibility

To *Numeracy* readers, the notion that results of the PIAAC numeracy assessment invite cause for concern may not come as a surprise. Scholars in our community have taken great strides to develop and report on assessments that invite more than just the capacity to correctly answer multiple-choice or fill-in-the-blank questions, the rationale being that alternative assessments might "show whether students have strengthened a tendency to use that capacity or have developed the skills necessary to deploy the capacity effectively in contexts other than those in the test" (Grawe et al. 2010, 1). Though not specifically grounded in the language of a social practices approach to numeracy, such work—in congruence with that approach—highlights the notion that if we seek to understand what students do (i.e., their practices), we should provide them with the freedom and space to tell us what it is that they do. If the assessments we use to elicit what students do sacrifice that space to account for constraints such as time, efficiency, or culture, then it is imperative that we acknowledge that sacrifice and qualify our work appropriately. In light of the apparent limitations of large-scale assessments to capture nuance in what individuals do with numeracy, a separate and new line of research might endeavor to understand what it is that we can learn from large-scale assessments. Indeed, it is likely that there are claims that we can make about individuals' numeracy practices based on the numeracy events they engage in as part of an assessment; however, it would require *new* and nontrivial work to make these connections.

As scholars of numeracy, we know all too well that data is subject to interpretation. The ways that we report our work are informed by a series of decisions that we make, whether conscious or unconscious, and ultimately those decisions influence how our work might be taken up by others. Just as we desire for our students (Polito 2014), or for journalists (Yarnall and Ranney 2017), to be aware of how quantitative information can be communicated, so too should we take it upon ourselves to consider how the information *we* communicate more broadly can be communicated. In the context of the PIAAC numeracy assessment, I have argued that nontrivial lapses in communication suggest that the assessment measures something that it does not. We should be aware of these lapses by interrogating statistics about test scores, by carefully hedging the ways that we talk about large-scale assessments (Evans 2014), and by—as responsible consumers and

producers of information—seeking out more information before assuming we have the full story.

Beyond what may seem trite or obvious to some, I hope this analysis has provided information for scholars to consider in developing their numeracy assessments in the future. In particular, I have outlined sources of evidence to consider in making judgments about validity for an assessment (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014), including those pertaining to a test's content, its internal structure, the ways test-takers are to respond, relationships among the variables it aims to measure, and its consequences. Though not all of these sources may be necessary for supporting an interpretation with a given use in mind—especially when the scope or consequences of one's assessment may be smaller than those of PIAAC—it is imperative that one be aware of where experts in assessment validation currently stand (Cizek et al. 2008). Awareness of existing scholarship is critical to developing a robust collective literature base around numeracy (Scheaffer 2008), even as our individual understandings and work vary in epistemology, method, and purpose.

Acknowledgments

I am grateful to my dissertation committee of Tonya Bartell, Lynn Fendler, Beth Herbel-Eisenmann, and Vince Melfi, as well as the Reviewers and Editors of *Numeracy* for their constructive feedback in developing this manuscript.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Baker, Dave, and Brian Street. 1994. "Literacy and Numeracy: Concepts and Definitions." In *Encyclopaedia of Education 1994*, edited by Husén, Torsten, and Neville Postlethwaite. Oxford: Pergamon Press.

Cahoon, Andrew, and Semra Kiliç-Bahi. 2019. "Assessing Quantitative Literacy: Challenges and Opportunities." In *Shifting Contexts, Stable Core: Advancing Quantitative Literacy in Higher Education*, edited by Tunstall, Samuel Luke, Gizem Karaali, and Victor Piercey, 185–96. Washington, DC: Mathematical Association of America.

Carraher, Terezinha Nunes, David William Carraher, and Analúcia Dias Schliemann. 1985. "Mathematics in the Streets and in Schools." *British*

Journal of Developmental Psychology 3(1): 21–9. https://doi.org/10.1111/j.2044-835X.1985.tb00951.x.

Cizek, Gregory J., Sharyn L. Rosenberg, and Heather H. Koons. 2008. "Sources of Validity Evidence for Educational and Psychological Tests." *Educational and Psychological Measurement* 68(3): 397–412. https://doi.org/10.1177/0013164407310130.

Clark, Lee Anna, and David Watson. 1995. "Constructing Validity: Basic Issues in Objective Scale Development." *Psychological Assessment* 7(3): 309. https://doi.org/10.1037/1040-3590.7.3.309.

Craig, Jeffrey, and Lynette Guzmán. 2018. "Six Propositions of a Social Theory of Numeracy: Interpreting an Influential Theory of Literacy." *Numeracy* 11(2): Article 1. https://doi.org/10.5038/1936-4660.11.2.2.

Emanuel, Gabrielle. 2016. "America's High School Graduates Look Like Other Countries' High School Dropouts." *NPR*, March 10. https://www.npr.org/sections/ed/2016/03/10/469831485/americas-high-school-graduates-look-like-other-countries-high-school-dropouts.

Evans, Jeff. 2014. "New PIAAC Results: Care Is Needed in Reading Reports of International Surveys." *Adults Learning Mathematics: An International Journal* 9(1): 37–52

Foote, Mary Q., and Tonya Gau Bartell. 2011. "Pathways to Equity in Mathematics Education: How Life Experiences Impact Researcher Positionality." *Educational Studies in Mathematics* 78(1): 45–68. https://doi.org/10.1007/s10649-011-9309-2.

Foucault, Michel. 1980. *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977*. First American ed. New York: Pantheon Books.

Frankel, Todd C. 2015. "U.S. Millennials Post 'Abysmal' Scores in Tech Skills Test, Lag behind Foreign Peers" *The Washington Post*, March 2. https://www.washingtonpost.com/news/wonk/wp/2015/03/02/u-s-millennials-post-abysmal-scores-in-tech-skills-test-lag-behind-foreign-peers/?utm_term=.b787a5c59ade.

Gal, Iddo, and Dave Tout. 2014. "Comparison of PIAAC and PISA Frameworks for Numeracy and Mathematical Literacy." *OECD Education Working Papers*, 102.

Gaze, Eric, Aaron Montgomery, Semra Kiliç-Bahi, Deann Leoni, Linda Misener, and Corrine Taylor. 2014. "Towards Developing a Quantitative Literacy/Reasoning Assessment Instrument." *Numeracy* 7(2): Article 4. https://doi.org/10.5038/1936-4660.7.2.4.

Gergen, Mary, and Kenneth Gergen. 2000. "Qualitative Inquiry: Tensions and Transformations." In *Handbook of Qualitative Research*, 2nd ed., edited by Denzin, Norman, and Yvonna Lincoln, 1025–46.Thousand Oaks, CA: Sage.

Graff, Harvey J. 1978. *The Literacy Myth: Literacy and Social Structure in the Nineteenth Century City*. New York: Academic Press.

Grawe, Nathan D., Neil S. Lutsky, and Christopher J. Tassava. 2010. "A Rubric for Assessing Quantitative Reasoning in Written Arguments." *Numeracy* 3(1): Article 3. https://doi.org/10.5038/1936-4660.3.1.3.

Grawe, Nathan D. 2011. "Beyond Math skills: Measuring Quantitative Reasoning in Context." *New Directions for Institutional Research* 149: 41–52. https://doi.org/10.1002/ir.379.

Griffin, Patrick, Esther Care, and Mark Wilson, eds. 2018. *Assessment and Teaching of 21st Century Skills: Methods and Approach*. New York: Springer.

Kahan, Daniel M., Ellen Peters, Erica C. Dawson, and Paul Slovic. 2017. "Motivated Numeracy and Enlightened Self-government." *Behavioural Public Policy* 1(1): 54–86. https://doi.org/10.1017/bpp.2016.2.

Kane, Michael. 2012. "All Validity Is Construct Validity. Or Is It?" *Measurement: Interdisciplinary Research & Perspective* 10(1–2): 66–70. https://doi.org/10.1080/15366367.2012.681977.

Karaali, Gizem, Edwin Villafane-Hernandez, and Jeremy Taylor. 2016. "What's in a Name? A Critical Review of Definitions of Quantitative Literacy, Numeracy, and Quantitative Reasoning." *Numeracy* 9(1): Article 2. https://doi.org/10.5038/1936-4660.9.1.2.

Klein, Stephen, Roger Benjamin, Richard Shavelson, and Roger Bolus. 2007. "The Collegiate Learning Assessment: Facts and Fantasies." *Evaluation Review* 31(5): 415–39. https://doi.org/10.1177/0193841X07303318.

Kline, Rex B. 2015. *Principles and Practice of Structural Equation Modeling*. New York: Guilford Publications.

Kosko, Karl W., and Jesse L. Wilkins. 2011. "Communicating Quantitative Literacy: An Examination of Open-ended Assessment Items in TIMSS, NALS, IALS, and PISA." *Numeracy* 4(2): Article 3. https://doi.org/10.5038/1936-4660.4.2.3.

Lather, Patti. 1993. "Fertile Obsession: Validity After Poststructuralism." *The Sociological Quarterly 34*(4): 673–93. https://doi.org/10.1111/j.1533-8525.1993.tb00112.x.

Lave, Jean, and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511815355.

Messick, Samuel. 1989. "Validity." In *Educational Measurement*, 3rd ed., edited by Robert L. Linn, 13–103. New York: Macmillan.

Mislevy, Robert J., and Geneva D. Haertel. 2006. "Implications of Evidence-Centered Design for Educational Testing." *Educational Measurement: Issues and Practice* 25(4): 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x.

National Commission on Excellence in Education. 1983. *A Nation at Risk: The Imperative for Educational Reform.* Washington, DC: National Commission on Excellence in Education.

Newton, Paul E. 2012. "Clarifying the Consensus Definition of Validity." *Measurement: Interdisciplinary Research & Perspective* 10(1–2): 1–29. https://doi.org/10.1080/15366367.2012.669666.

Newton, Paul E., and Jo-Anne Baird. 2016. "The Great Validity Debate." *Assessment in Education: Principles, Policy & Practice* 23(2): 173–7. https://doi.org/10.1080/0969594X.2016.1172871.

Organisation for Economic Co-operation and Development (OECD). 2012. *Literacy, Numeracy and Problem Solving in Technology-rich Environments: Framework for the OECD Survey of Adult skills*. Paris: OECD Publishing.

OECD. 2013a. *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. Paris: OECD Publishing.

OECD. 2013b. *Skilled for Life?: Key Findings from the Survey of Adult Skills*. Paris: OECD Publishing.

OECD. 2013c. *Time for the U.S. to Reskill? What the Survey of Adult Skills Says*. Paris: OECD Publishing.

OECD. 2016a. *Technical Report of the Survey of Adult Skills (PIAAC)*. Paris: OECD Publishing.

OECD. 2016b. *The Survey of Adult Skills: Reader's Companion*, 2nd ed. Paris: OECD Publishing.

Osborne, Jason W. 2008. "Is Disattenuation of Effects a Best Practice?" In *Best Practices in Quantitative Methods*, edited by Jason W. Osborne, 239–245. Thousand Oaks, CA: SAGE Publications, Inc. https://doi.org/10.4135/9781412995627.d20.

Oughton, Helen M. 2009. "A Willing Suspension of Disbelief? 'Contexts' and Recontextualisation in Adult Numeracy Classrooms." *Adults Learning Mathematics: An International Journal* 4(1): 16–31.

Oughton, Helen M. 2018. "Disrupting Dominant Discourses: A (Re) Introduction to Social Practice Theories of Adult Numeracy." *Numeracy* 11(1): Article 2. https://doi.org/10.5038/1936-4660.11.1.2.

PIAAC Numeracy Expert Group. 2009. "PIAAC Numeracy: A Conceptual Framework." *OECD Education Working Papers* 35.

Polito, Jessica. 2014. "The Language of Comparisons: Communicating about Percentages." *Numeracy* 7(1): Article 6. https://doi.org/10.5038/1936-4660.7.1.6.

Rear, David. 2019. "One Size Fits All? The Limitations of Standardised Assessment in Critical Thinking." *Assessment & Evaluation in Higher Education* 44(5): 664–675. https://doi.org/10.1080/02602938.2018.1526255.

Rhodes, Terrel L, ed. 2010. *Assessing Outcomes and Improving Achievement: Tips and Tools for Using Rubrics*. Washington, DC: Association of American Colleges and Universities.

Rice, Mark. 2009. "On Education, The U.S. Doesn't Measure Up." *Forbes*, October 22. https://www.forbes.com/2009/10/22/public-education-funding-oecd-opinions-contributors-mark-rice.html#6548d2e21f26.

Roohr, Katrina, HyeSun Lee, Jun Xu, Ou Liu, and Zhen Wang. 2017. "Preliminary Evaluation of the Psychometric Quality of HEIghten™ Quantitative Literacy." *Numeracy* 10(2): Article 3. https://doi.org/10.5038/1936-4660.10.2.3.

Scheaffer, Richard L. 2008. "Scientifically Based Research in Quantitative Literacy: Guidelines for Building a Knowledge Base." *Numeracy* 1(1): Article 3. https://doi.org/10.5038/1936-4660.1.1.3.

Scribner, Sylvia, and Michael Cole. 1981. *The Psychology of Literacy*. Cambridge, MA: Harvard University Press.

Shavelson, Richard J., Julián P. Mariño von Hildebrand, Olga Zlatkin-Troitschanskaia, and Susanne Schmidt. 2019. "Reflections on the Assessment of Quantitative Reasoning." In *Shifting Contexts, Stable Core: Advancing Quantitative Literacy in Higher Education*, edited by Tunstall, Samuel Luke, Gizem Karaali, and Victor Piercey, 163–76. Washington, DC: Mathematical Association of America.

Shulman, Lee S. 1981. "Disciplines of Inquiry in Education: An Overview." *Educational Researcher* 10(6): 5–23. https://doi.org/10.3102/0013189X010006005.

Sireci, Stephen G., and Tia Sukin. 2013. "Test Validity." In *APA Handbook of Testing and Assessment in Psychology*, Volume 1, 61–84. Washington, DC: American Psychological Association. https://doi.org/10.1037/14047-004.

Steen, Lynn A., ed., and National Council on Education and the Disciplines (NCED). 2001. *Mathematics and Democracy: The Case for Quantitative Literacy*. Princeton, NJ: NCED.

Stephanopoulos, Nicholas O., and Eric M. McGhee. 2015. "Partisan Gerrymandering and the Efficiency Gap." U. Chi. L. Rev. 82(2): 831–900.

Terman, Lewis M., Grace Lyman, George Ordahl, Louise Ordahl, Neva Galbreath, and Wilford Talbert. 1915. "The Stanford Revision of the Binet-Simon Scale and Some Results from Its Application to 1000 Non-selected Children. *Journal of Educational Psychology* 6(9): 551–62. https://doi.org/10.1037/h0075455.

Thorndike, Edward L. 1916. *An Introduction to the Theory of Mental and Social Measurements*, 2nd ed. New York, NY: Teachers College, Columbia University.

Tout, Dave, Diana Coben, Vince Geiger, Linda Ginsburg, Kess Hoogland, Terry Maguire, Sue Thompson, and Ross Turner. 2017. *Review of the PIAAC Numeracy Assessment Framework: Final Report.* Camberwell, Australia: Australian Council for Educational Research.

Tunstall, Samuel L., Rebecca L. Matz, and Jeffrey C. Craig. 2018. "Quantitative Literacy Courses as a Space for Fusing Literacies." *The Journal of General Education* 65(3–4): 178–94. https://doi.org/10.5325/jgeneeduc.65.3-4.0178.

Tunstall, Samuel L., Vincent Melfi, Jeffrey C. Craig, Richard Edwards, Andrew Krause, Bronlyn Wassink, and Victor Piercey. 2016. "Quantitative Literacy at Michigan State University, 3: Designing General Education Mathematics Courses." *Numeracy* 9(2): Article 6. https://doi.org/10.5038/1936-4660.9.2.6.

Vacher, H. L. 2015. "Educational Assessment Is an Enduring Theme of *Numeracy*." *Numeracy* 8(1): Article 1. https://doi.org/10.5038/1936-4660.8.1.1.

Wolcott, Harry F. 1990. "On Seeking—and Rejecting—Validity in Qualitative Research." In *Qualitative Inquiry in Education: Continuing the Debate*, edited by Eisner, Elliot, and Alan Peshkin, 121–52. New York, NY: Teachers College Press.

Woloshin, Steven, and Lisa M. Schwartz. 2002. "Press Releases: Translating Research into News." *JAMA* 287(21): 2856–8. https://doi.org/10.1001/jama.287.21.2856.

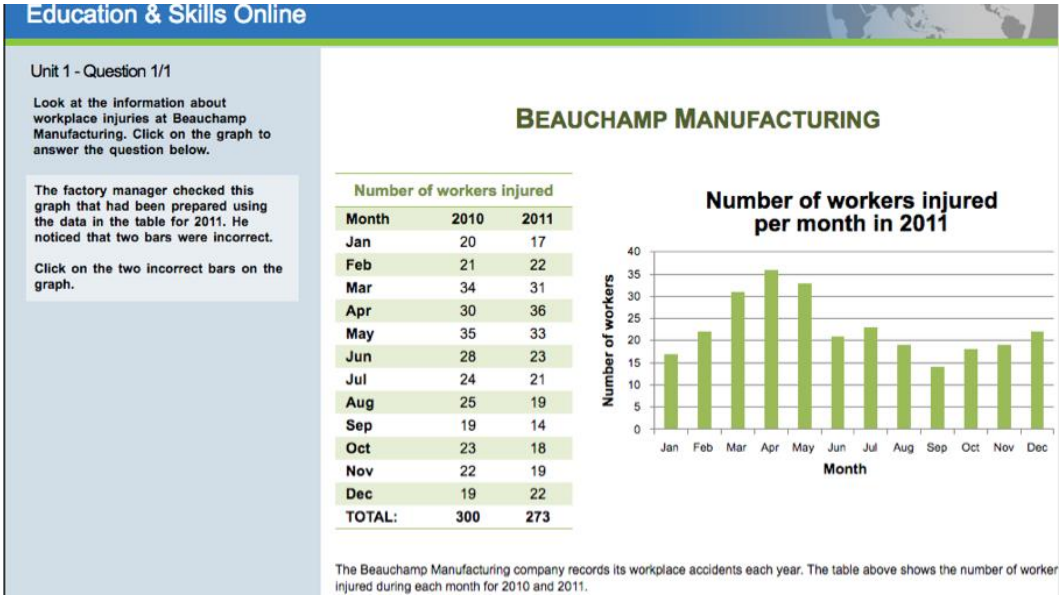
Yarnall, Louise, and Michael Andrew Ranney. 2017. "Fostering Scientific and Numerate Practices in Journalism to Support Rapid Public Learning." *Numeracy* 10(1): Article 3. https://doi.org/10.5038/1936-4660.10.1.3.

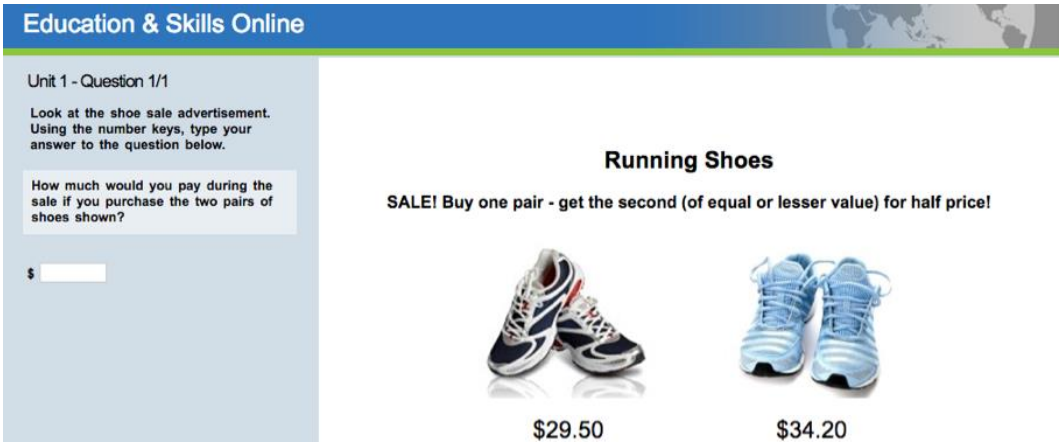
Zerr, Ryan. 2019. "Assessing Quantitative Literacy as a Cumulatively-Acquired Intellectual Skill." In *Shifting Contexts, Stable Core: Advancing Quantitative Literacy in Higher Education*, edited by Tunstall, Samuel Luke, Gizem Karaali, and Victor Piercey, 177–84. Washington, DC: Mathematical Association of America.

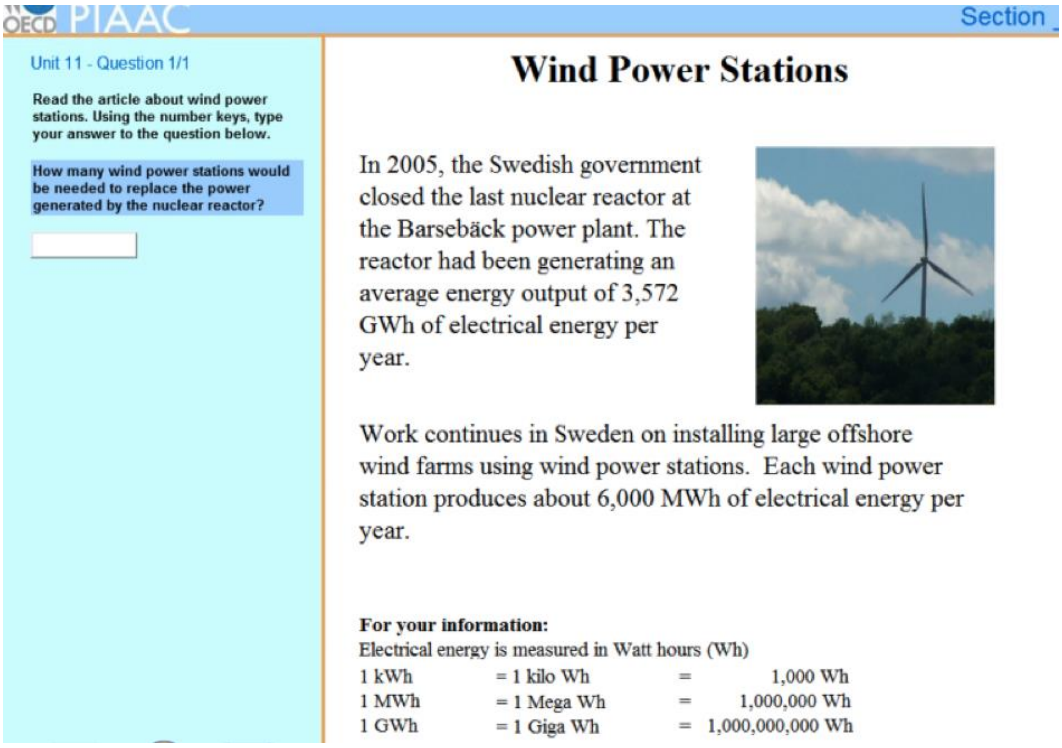
Zinshteyn, Mikhail. 2015. "The Skills Gap: America's Young Workers Are Lagging Behind." *The Atlantic*, February 17. https://www.theatlantic.com/education/archive/2015/02/the-skills-gap-americas-young-workers-are-lagging-behind/385560/.

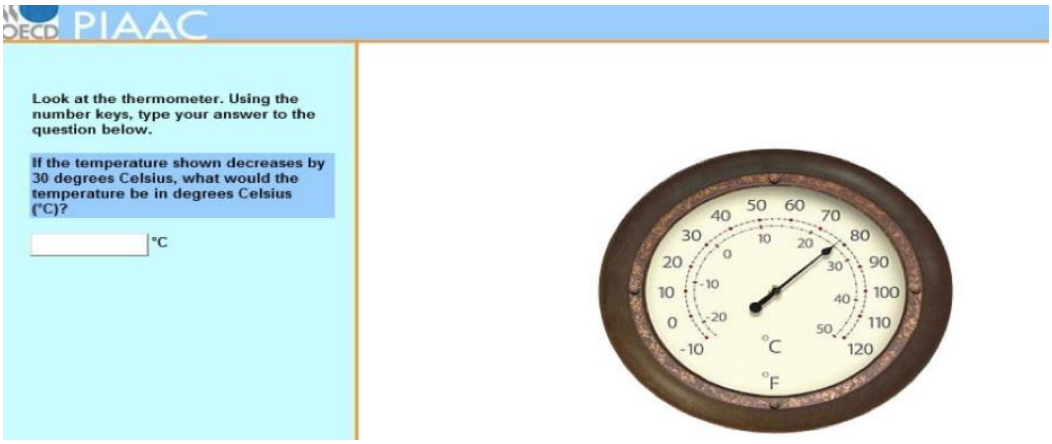
Appendix

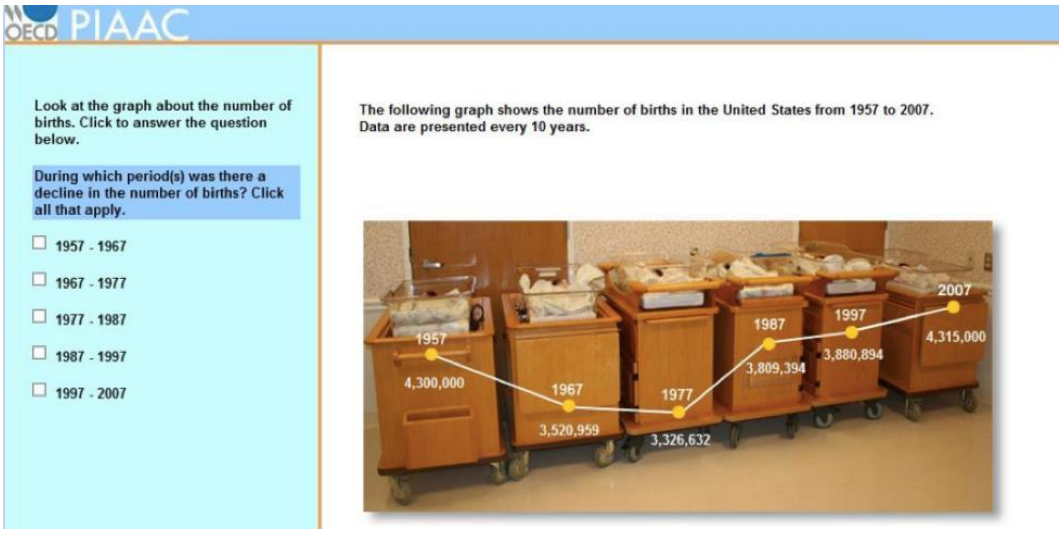
Five publicly available items from the OECD's PIAAC numeracy assessment (See the sample test environment at http://www.oecd.org/skills/ESonline-assessment/takethetest/? for the first two items, and http://www.oecd.org/skills/piaac/Numeracy%20Sample%20Items.pdf for the last three items)






PIAAC
Section
Unit 11 - Question 1/1
Read the article about wind power stations. Using the number keys, type your answer to the question below.
How many wind power stations would be needed to replace the power generated by the nuclear reactor?
Wind Power Stations
In 2005, the Swedish government closed the last nuclear reactor at the Barsebäck power plant. The reactor had been generating an average energy output of 3,572 GWh of electrical energy per year.
Work continues in Sweden on installing large offshore wind farms using wind power stations. Each wind power station produces about 6,000 MWh of electrical energy per year.
For your information:
Electrical energy is measured in Watt hours (Wh)
1 kWh = 1 kilo Wh = 1,000 Wh
1 MWh = 1 Mega Wh = 1,000,000 Wh
1 GWh = 1 Giga Wh = 1,000,000,000 Wh


PIAAC
Look at the thermometer. Using the number keys, type your answer to the question below.
If the temperature shown decreases by 30 degrees Celsius, what would the temperature be in degrees Celsius (°C)?
°C
°C
°F


PIAAC
Look at the graph about the number of births. Click to answer the question below.
During which period(s) was there a decline in the number of births? Click all that apply.
1957 - 1967
1967 - 1977
1977 - 1987
1987 - 1997
1997 - 2007
The following graph shows the number of births in the United States from 1957 to 2007. Data are presented every 10 years.
1957
4,300,000
1967
3,520,959
1977
3,326,632
1987
3,809,394
1997
3,880,894
2007
4,315,000